Statistical Data Analysis

In order to properly model my data, I first compiled all of the necessary information into a single dataframe that contained the amount each team spent per year, along with a breakdown on how much each team spent per age group and position group. I also included the position value, which assigns a value to a team based on league position. A team that finished in 1st place in its league will have a position value of 100, a 2nd place finish would yield a position value of 95, and so on.

I first tested the correlation between all variables with a correlation matrix, as well as two OLS linear regression models with position value as the dependent variable (Output 4, 6, & 7). I found that there is not a strong direct correlation between the position value and any of the spending categories, with R-squared values for these models at around 0.1. There is, however, a strong correlation between position value and the performance categories, such as goals scored, goals against, and goal difference. This is to be expected, as it clearly follows that if a team scores more and concedes less, they will win more games and finish higher in the league table.

Because of this, I decided to look at the relationship between the performance categories and spending categories. Goal difference is highly correlated to position value (correlation coefficient of 0.84), so I generated OLS models with goal difference as the dependent variable and the spending categories as the independent variables (Output 8 & 9). The R-squared values for these models were slightly higher than the previous set of models at almost 0.2, but these would still not suggest a strong correlation between spending and results. Despite the low R-squared values for these models, with the exception of the 30+ category, the p-values for each variable were very low, so I would reject the null hypothesis that these variables have no effect on the dependent variable.

One reason for the low R-squared values in these models could be that it takes more than a year for a team to see an improvement in results after spending a significant amount of money, so I created two additional columns in the dataframe. The first is a two year total amount spent (null for the first year of data for each team), and the second is a three year total amount spent (null for the first two years of data for each team. I generated two OLS models, both with goal difference as the dependent variable, one with two year total as the independent variable, and one with three year total as the independent variable (Output 10 & 11). The R-squared values for the two and three year models were 0.23 and 0.27, respectively, so while it did slightly increase from the previous models, it was not by a significant amount.

The last method I tried to account for the time it can take for money spent to affect results was to take the average over the time period of data collected. I grouped the original dataframe by team

and took the average of each category to create a dataframe with a single row per team and columns containing the average values of all of the original categories. I generated an OLS model with average goal difference as the dependent variable and all of the spending categories as well as average age as the independent variables (Output 13). With an R-squared of 0.52, this model explains significantly more of the variance in goal difference than any of the previous models, while the most significant variables look to be midfielders, players between 15 and 22, and the average age of players bought.

Many top teams are known to have extraordinary youth academies where they develop young players and integrate them into the full team at no cost to the club, which could account for why some teams have achieved good results without spending much money. Another reason is that many teams take advantage of free transfers, which are players who are out of contract and do not have a parent club that they need to pay in order to sign them to a contract.